PENINGKATAN INTERLINKING PADA LINKED DATA HETEROGEN MELALUI ONTOLOGY ALIGNMENT

Inne Gartina Husein[1)], **Benhard Sitohang**[2)], **Saiful Akbar**[3)]

[1), 2), 3)] *Sekolah Tinggi Elektro dan Informatika Institut Teknologi Bandung*
Jalan Ganesa No. 10, Bandung 40132
[1)] *Fakultas Ilmu Terapan Universitas Telkom*
Jalan Telekomunikasi No.1 Terusan Buah Batu, Bandung 40257
Email : innegh235@students.itb.ac.id [1)], benhard@informatika.org [2)], saiful@informatika.org [3)]

Abstrak

Membangun interlinking pada Web of data merupakan salah satu isu penting, agar jaringan data dari berbagai sumber data heterogen dapat saling terhubung. Berbagai macam cara telah dilakukan para pengembang, seperti cara manual, transformasi URI dan memanfaatkan satu ontologi. Demikian juga telah dikembangkan beberapa tool untuk membangun link. Namun tidak ada diantara cara dan tool yang telah dikembangkan memanfaatkan ontology alignment (OA) di dalam pencarian relasi antar entitas pada sumber data terutama untuk topik data yang berbeda. Tujuan makalah ini adalah menelaah hubungan antara OA dengan membangun interlinking pada Web of data. Hasilnya adalah berupa keterkaitan yang penting diantara keduanya, yakni dimana OA dapat dimanfaatkan untuk menemukan relasi antar entitas sehingga memudahkan pengembang (ataupun tool) dalam membangun interlinking yang kuat.

Kata kunci : *interlinking, Web of data, ontology alignment, relasi, entitas*

1. Pendahuluan

Web of data merupakan sebuah jaringan yang dihasikan dari sumber data terstruktur yang dipublikasikan ke Web. Kumpulan data dari berbagai sumber tersebut dihubungkan dengan link yang dinyatakan secara eksplisit, yang disebut Linked Data (LD).

Salah satu isu penting pada LD adalah bagaimana membangun keterhubungan antar berbagai kumpulan data yang heterogen secara semantik. Heterogenitas secara semantik dapat dijelaskan dari sisi terminologi, yakni pada saat dua istilah berbeda digunakan untuk merepresentasikan entitas yang sama pada dua buah kumpulan data. Sebagai contoh adalah paper vs articulo, atau paper vs memo, atau paper vs article [1]. Istilah adalah kosa kata (*vocabulary*) yang digunakan untuk merepresentasikan entitas-entitas pada kumpulan data.

Kumpulan data pada *Web of data* direpresentasikan dengan satu atau lebih *vocabulary* atau ontologi, mulai dari skema basis data sederhana sampai ontologi berskala besar [2].

Pada sisi lain dikatakan bahwa terdapat lebih dari 31 triliun pernyataan RDF yang dipublikasikan sebagai LD namun hanya terdapat 500 juta link di dalamnya. Hal ini menunjukkan lemahnya keterhubungan dari satu kumpulan data ke kumpulan data lainnya [3].

Manfaat utama dari keterhubungan tersebut adalah terciptanya integrasi data, yang memungkinkan terjadinya pertukaran dan bahkan berbagi data pada *Web of data*. Dengan demikian dibutuhkan suatu cara untuk menemukan keterhubungan atau link pada berbagai *vocabulary* atau ontologi tersebut yang dapat meningkatkan keterhubungan pada link data, baik dalam jumlah maupun kualitas link.

Tujuan dari makalah ini adalah mempertimbangkan kegiatan pensejajaran ontologi (*ontology alignment*) dan kegiatan membangun keterhubungan (*interlinking*) pada LD sebagai dua kegiatan yang saling terkait. Serta bagaimana kedua kegiatan tersebut dapat saling mendukung dalam meningkatkan *interlinking* pada LD.

Setelah menjelaskan mengenai latar belakang, makalah ini akan membahas mengenai berbagai pendekatan yang telah dilakukan sebelumnya berkaitan dengan peningkatan *interlinking*, kemudian membahas prinsip-prinsip LD, dan *ontology alignment*. Bagian terakhir adalah kesimpulan yang dihasilkan.

2. Pembahasan

Pada bagian ini dibahas beberapa pendekatan yang telah dilakukan guna membangun *interlinking*.

A. ODD-linker

Sebuah *interlinking tool* yang digunakan untuk mencari record yang sama dengan teknik mining. ODD-linker menggunakan query SQL untuk mengidentifikasi dan membandingkan *duplicate record* pada basis data relasional. Keluaran dari sistem ini adalah spesifikasi link dalam bahasa LinQL yang menggambarkan *duplicate record* yang menjadi acuan untuk membangun *interlinking* pada sumber data yang akan dipublikasikan sebagai LD [4]. Teknik *interlinking* tersebut mengandalkan pada record yang sama namanya dan berulang pada dua kumpulan data relasional.

B. R2R Vocabulary Mapping

Sebuah framework yang dibangun untuk memetakan korespondensi istilah dengan cara **mengubah** istilah yang digunakan sumber data asal (*source*) menjadi istilah yang sesuai dengan sumber data yang dituju (*target*). Framework ini mendukung keterlibatan

pengguna dalam proses pemetaan *vocabulary* sehingga pengguna dapat melakukan edit dan eksekusi proses pemetaan [3]. Dengan kata lain, proses pengubahan istilah dilakukan secara manual oleh pengguna, sehingga keberhasilan proses mapping sangat tergantung oleh ketelitian dan penguasaan ilmu pengguna.

C. Silk - Link Discovery Framework

Silk adalah sebuah framework yang menghasilkan spesifikasi link guna menggambarkan hubungan semantik antar dua entitas. Spesifikasi link dibuat secara manual oleh pengguna melalui antarmuka Silk atau dapat juga menggunakan spesifikasi link yang telah dibuat sebelumnya oleh orang lain. Berdasarkan spesifikasi link yang telah dibuat sebelumnya, maka akan dibuat link secara otomatis [3], [5].

Masalah Pada Interlinking

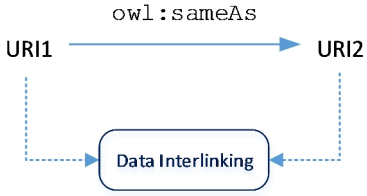
Beberapa prinsip mengenai *Web of data* menurut Tim Berners-Lee [6] yakni :

1) Menggunakan URI sebagai nama sesuatu
2) Menggunakan HTTP URI sehingga pengguna dapat melihat nama tersebut
3) Menggunakan standar RDF dan SPARQL dalam mencari URI dan menyediakan informasi berguna
4) Terdapat link ke URI lainnya, sehingga pengguna dapat menemukan hal-hal lain.

Masalah utama pada *Web of data* adalah menciptakan link antar entitas-entitas yang ada pada kumpulan data yang berbeda. Seperti disebutkan di atas bahwa link merupakan hubungan antar URI pada LD. Beberapa skema URI yang dapat digunakan adalah [7]:

- `ftp://ftp.is.co.za/rfc/rfc1808.txt`
- `http://www.ietf.org/rfc/rfc2396.txt`
- `ldap://[2001:db8::7]/c=GB?objectClass?one`
- `mailto:John.Doe@example.com`
- `news:comp.infosystems.www.servers.unix`
- `tel:+1-816-555-1212`
- `telnet://192.0.2.16:80/`

Umumnya link merupakan peryataan `owl:sameAs`, maka ilustrasi link pada LD dapat dilihat pada gambar 1 di bawah ini [2]. Saat diidentifikasi adanya link antar kumpulan data, maka link tersebut harus dipublikasikan agar dapat digunakan ulang oleh kumpulan data lainnya.

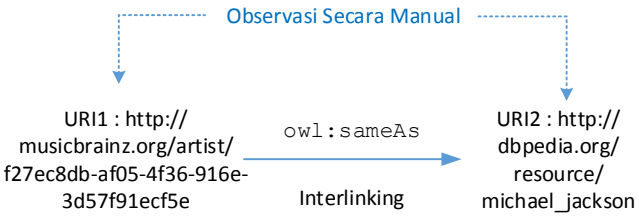


Gambar 1 Link dengan `owl:sameAs`

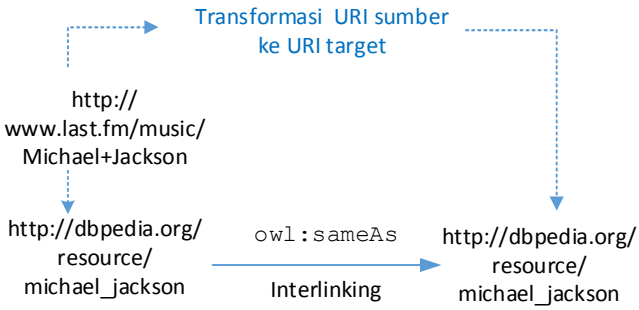
Berbagai cara dapat dilakukan untuk mengidentifikasikan link antar kumpulan data, seperti cara manual, cara transformasi URI dan cara membandingkan dengan sebuah ontologi [2]. Cara manual adalah melakukan observasi manual ke sumber data dan dengan bantuan suatu tool maka dibuatlah link antar URI yang dianggap sama (gambar 2).

Suatu aturan dapat dibuat untuk mengidentifikasi kesamaan sumber data, misalnya penamaan seorang artis pada kumpulan data. Sebagai contoh kumpulan data Last.FM merepresentasi URI untuk seorang artis adalah dengan pola "nama_depan + nama_belakang". Sedangkan pada Dbpedia menggunakan pola "namadepan_namabelakang". Penamaan dapat diseragamkan mengikuti URI target. Sebagai ilustrasi dapat dilihat pada gambar 3.

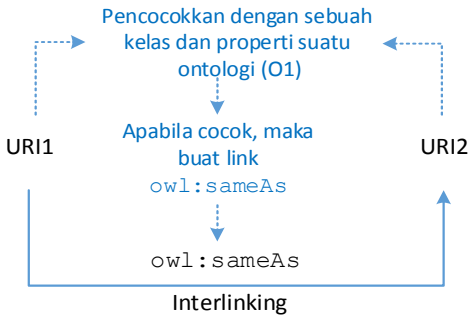
Kedua cara (gambar 2 dan gambar 3) masih menggunakan cara-cara manual dan belum memanfaatkan *vocabulary* atau ontologi dalam membangun *interlinking*.



Gambar 2 Manual *Interlinking* [8]



Gambar 3 Transformasi URI

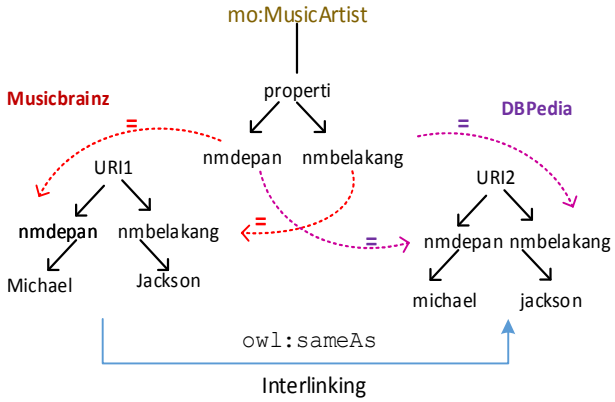


Gambar 4 Mencocokkan dengan sebuah ontologi

Diketahui bahwa sumber data pada LD merepresentasikan data ke dalam *vocabulary* atau ontologi. Bagaimana hubungan antara URI dengan ontologi, baik URI sumber maupun target akan dicocokkan dengan sebuah ontologi. Cara ketiga

merupakan proses identifikasi entitas dengan mencocokannya dengan sebuah kelas dan properti pada ontologi tersebut (gambar 4).

Sebagai contoh, URI1 dan URI2 dicocokkan dengan sebuah kelas pada sebuah Music Ontology (MO). Apabila kedua URI memiliki kesamaan dengan properti dari kelas (pada ontologi) maka kedua URI didefinisikan memiliki hubungan atau link. Ilustrasi dapat dilihat pada gambar 5.



Gambar 5 Ilustrasi Pencocokkan kelas MusicArtist

Ketiga cara yang telah dijelaskan di atas adalah cara-cara pembangunan *interlinking* pada level *instance*, yaitu pada URI, dan belum menyentuh level diatasnya yakni level skema.

Pemanfaatan ontologi sebagai sumber daya pengetahuan pada LD masih sebatas mencocokan dengan kelas dan/atau properti yang sama. Apabila pembangunan *interlinking* pada LD hanya berpaku pada level *instance* maka akan menghasilkan hubungan yang lemah atau disebut *very loosely coupled* [9].

Web of data terdiri dari berbagai sumber data yang heterogen, baik dari sisi semantik maupun dari sisi domain pengetahuan. Heterogenitas semantik antara lain adalah perbedaan penggunaan terminologi pada berbagai sumber data. Hal ini terjadi saat dua istilah berbeda digunakan untuk merepresentasikan entitas yang sama. Sebagai contoh adalah paper vs articulo, atau paper vs memo, atau paper vs article. Heterogenitas dapat pula berarti perbedaan domain pengetahuan artinya setiap sumber data berbeda-beda topiknya. Sebagai contoh topik pemerintahan, hiburan (musik, film, artis), bio-informatika, publikasi, geospasial dan yang lainnya.

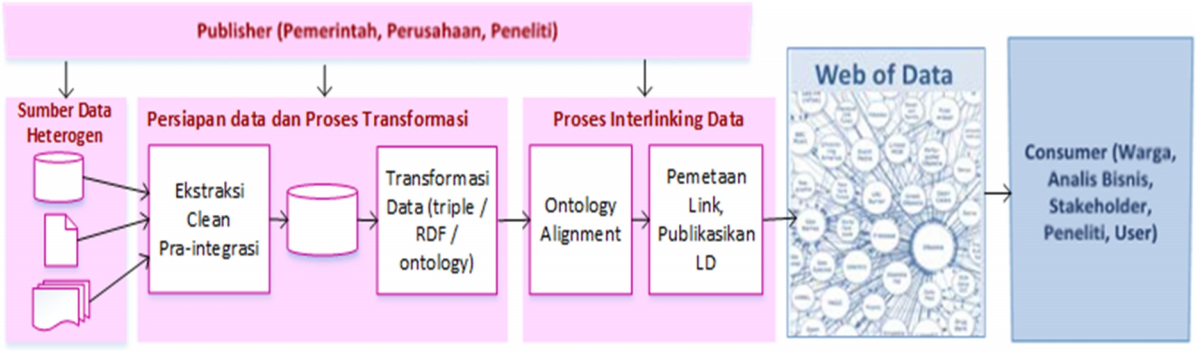
Maka dengan itu dibutuhkan proses *interlinking* yang dapat menghubungkan berbagai sumber data baik secara semantik maupun secara domain pengetahuan [10].

Ontology Alignment Mendukung ProsesInterlinking

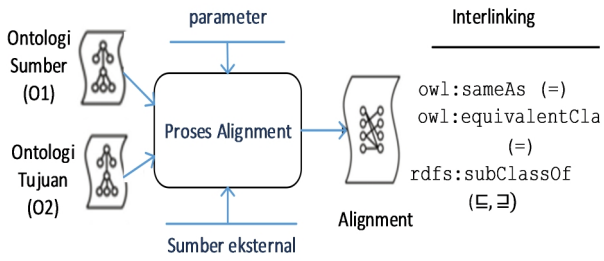
Pensejajaran ontologi atau *ontology alignment* adalah identifikasi relasi antar elemen yang individual dari berbagai ontologi dengan tujuan membangun interoperabilitas antar sumber data yang menggunakan ontologi yang berbeda secara individual [11]. Tujuan dari *ontology alignment* (OA) adalah menghilangkan heterogenitas tersebut sehingga menghasilkan korespondensi pada dua atau lebih sumber data [12]. Output dari proses ini adalah sebuah *alignment* yang menggambarkan korespondensi antar dua buah ontologi yang dicocokkan.

Sebelum menjadi data terstruktur yang dipublikasikan, LD sendiri bersumber dari kumpulan data yang heterogen. Perlu dilakukan proses transformasi agar kumpulan data heterogen dapat diubah menjadi model data yang sama, yakni model data RDF. Setelah itu dibuatkan pemetaan relasi antar pernyataan RDF sehingga data dapat dipublikasikan pada Web sebagai data yang saling terhubung [6].

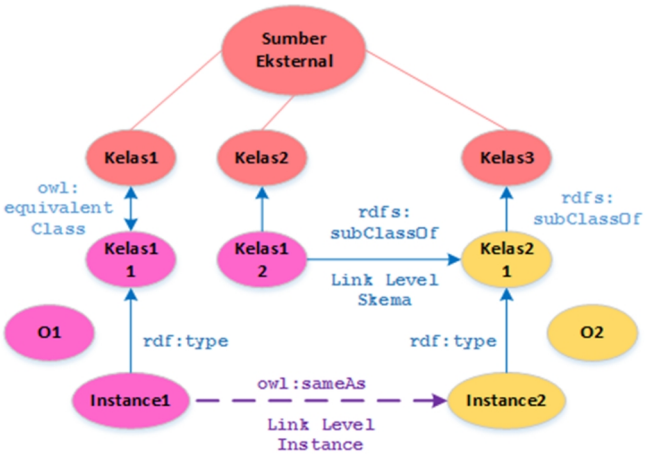
Terdapat kemiripan antara pemetaan relasi pada LD dengan *alignment* yang dihasilkan oleh proses OA, mengingat *alignment* adalah gambaran relasi antar elemen pada dua ontologi. Penulis mempertimbangkan untuk menyisipkan OA dalam proses interlinking pada LD. Sebagai ilustrasi dapat dilihat pada gambar 6 bahwasanya OA merupakan tahapan penting dalam membangun *interlinking* pada LD. OA membantu menemukan



Gambar 6 Proses *Interlinking* Data Menggunakan *Ontology Alignment*



Gambar 7 Proses Ontology Alignment [13]

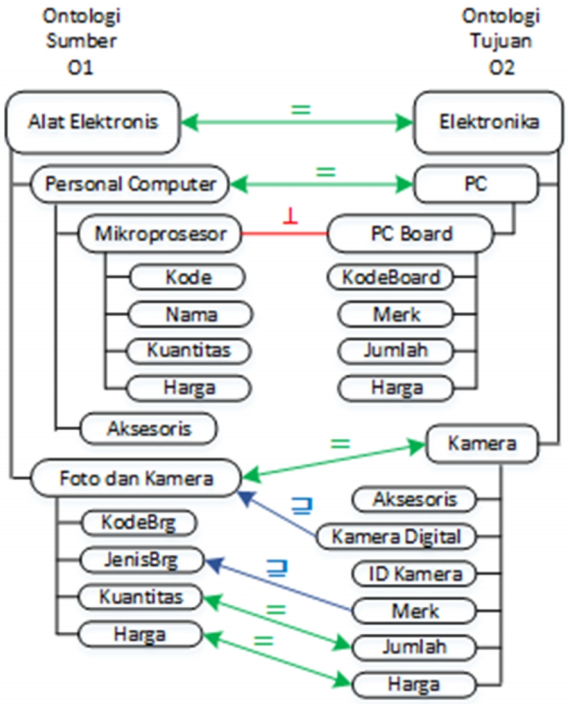


Gambar 8 Link Level Skema

Pada gambar 7 diilustrasikan bahwa korespondensi (atau alignment) antar entitas-entitas tidak hanya menghasilkan relasi ekivalen yaitu sama (=). Namun juga menghasilkan relasi disjoin atau tidak sama (⊥), serta menghasilkan relasi subsumption seperti lebih spesifik (⊑) dan lebih umum (⊒). Proses OA tidak hanya menghasilkan interlinking dengan level instance (yakni relasi owl:sameAs) namun juga level skema yakni relasi owl:equivalentClass dan rdfs:subClassOf. Link pada level skema akan menghasilkan keterhubungan yang kuat antar data pada LD [9].

Gambar 8 mengilustrasikan relasi pada level skema dengan memanfaatkan sumber eksternal. Diasumsikan kelas2 merupakan subkelas dari kelas3, sehingga menghasilkan relasi rdfs:subClassOf dari Kelas12 ke Kelas21.

Gambar 9 mengilustrasikan berbagai berbagai bentuk relasi yang dihasilkan dari OA, seperti sama (=), tidak sama (⊥), lebih spesifik (⊑) dan lebih umum (⊒) dari ontologi sumber O1 ke ontologi tujuan O2. Relasi yang dihasilkan menjadi acuan dalam membangun link pada level skema antar dua LD yang heterogen. Proses OA yang dilakukan secara otomatis oleh sistem dapat mengurangi kesalahan dalam membuat link apabila dibuat secara manual oleh pengguna, seperti pendekatan-pendekatan *interlinking* lain yang telah dibahas sebelumnya. Dengan demikian OA dapat meningkatkan validitas link pada LD sehingga mendukung integrasi data yang optimal.



Gambar 9 Ilustrasi Relasi Pada O1 dan O2 [14]

3. Kesimpulan

Salah satu isu penting pada LD adalah bagaimana membangun keterhubungan (atau *interlinking*) antar berbagai kumpulan data yang heterogen secara semantik. Diketahui terdapat lebih dari 31 triliun pernyataan RDF yang dipublikasikan sebagai LD namun hanya terdapat 500 juta link di dalamnya. Hal ini menunjukkan lemahnya keterhubungan dari satu kumpulan data ke kumpulan data lainnya [3].

Manfaat utama dari *interlinking* adalah terciptanya integrasi data, yang memungkinkan terjadinya pertukaran dan bahkan berbagi data pada *Web of data*. Dengan demikian dibutuhkan suatu cara untuk menemukan keterhubungan pada berbagai berbagai sumber data.

Tujuan dari penelitian ini adalah mempertimbangkan proses pensejajaran ontologi (*ontology alignment*) dan proses membangun keterhubungan (*interlinking*) pada LD sebagai dua kegiatan yang saling terkait. Serta bagaimana kedua kegiatan tersebut dapat saling mendukung dalam meningkatkan *interlinking* pada LD.

Berdasarkan hasil penelaah berbagai macam cara membangun interlinking yang ada dengan segala kekurangannya, maka *ontology alignment* (OA) menjadi satu proses yang dapat mengurangi kekurangan-kekurangan tersebut. OA menghasilkan relasi tidak hanya pada level instance namun juga pada level skema, yang **memperkuat** *interlinking* sehingga membantu dalam pertukaran dan berbagi data antar sumber data. Selain itu OA dapat melakukan *alignment* pada kelas yang bukan hanya relasi ekivalensi namun juga relasi subsumption pada level skema. Dimana diketahui bahwa link pada level skema lebih optimal dibandingkan link pada level instance. Pada gambar 6 dijelaskan bahwa OA merupakan tahapan penting pada proses *interlinking* data, dengan menyediakan korespondensi antar elemen secara otomatis. Dengan demikian OA dapat mendukung integrasi data pada *Web of data* secara optimal.

Daftar Pustaka

[1] J. Euzenat dan P. Shvaiko, Ontology Matching. Springer, 2013.
[2] F. Scharffe dan J. Euzenat, "Linked data meets ontology matching : enhancing data linking through ontology alignments," 2013.
[3] V. Bryl, C. Bizer, R. Isele, M. Verlic, S. G. Hong, S. Jang, M. Y. Yi, dan K. Choi, "Interlinking and Knowledge Fusion," in LNSC 8661, 2014, vol. 8661, pp. 70–89.
[4] O. Hassanzadeh, L. Lim, A. Kementsietsidis, dan M. Wang, "A declarative framework for semantic link discovery over relational data," in Proceedings of the 18th International Conference on World wide web, 2009, pp. 1101–1102.
[5] J. Volz, C. Bizer, M. Gaedke, dan G. Kobilarov, "Silk - A Link Discovery Framework for the Web of Data," in CEUR Workshop Proceedings, 2009, vol. 538.
[6] T. Berners-Lee, "Linked Data - Design Issues," 2009. [Online]. Available at: http://www.w3.org/DesignIssues/LinkedData.html. [Diakses: 30-Aug-2015].
[7] D. Wood, M. Zaidman, L. Ruth, dan M. Hauseblas, Linked Data : Structured Data on the Web. Manning Shelter Island, 2014.
[8] F. Scharffe dan J. Euzenat, "MeLinDa: an interlinking framework for the web of data," 2011.
[9] P. Jain, P. Z. Yeh, K. Verma, R. G. Vasquez, M. Damova, P. Hitzler, dan A. P. Sheth, "Contextual ontology alignment of LOD with an upper ontology: A case study with proton," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6643 LNCS, no. PART 1, pp. 80–92.
[10] P. Jain, P. Hitzler, A. A. P. Sheth, K. Verma, dan P. P. Z. Yeh, "Ontology alignment for linked open data," in Iswc, 2010, pp. 402–417.
[11] M. Ehrig, Ontology Alignment - Bridging the Semantic Gap. Springer, 2007.
[12] P. Shvaiko dan J. Euzenat, "Ontology matching: state of the art and future challenges," in Knowledge and Data Engineering, IEEE …, 2013, vol. 1, no. X, pp. 158–176.
[13] I. G. Husein, B. Sitohang, dan S. Akbar, "Ontology Matching : State of the Art," Bandung, Indonesia, 2014.
[14] P. Shvaiko dan J. Euzenat, Tutorial on Schema and Ontology Matching. 2005.

Biodata Penulis

Inne Gartina Husein, memperoleh gelar Sarjana Komputer (S.Kom.) Jurusan Manajemen Informatika STMIK LIKMI Bandung, lulus tahun 2000. Memperoleh gelar Magister Teknik (M.T.) Program Pasca Sarjana Magister Teknik Informatika Institut Teknologi Bandung, lulus tahun 2005. Saat ini terdaftar sebagai mahasiswa Program Doktoral Sekolah Teknik Elektro dan Informatika (STEI) Intitut Teknologi Bandung dan sebagai Dosen Tetap di Universitas Telkom Bandung.

Benhard Sitohang, memperoleh gelar Insinyur (Ir), Jurusan Teknik Elektro Institut Teknologi Bandung, lulus tahun 1978. Memperoleh gelar Magister Bidang Basis Data (DEA) dari Universite des Sciences et Techniques du languadoc, Perancis, lulus tahun 1980. Memperoleh gelar Doktor Bidang Basis Data (Dr.Ing.) dari Universite des Sciences et Techniques du languadoc, Perancis, lulus tahun 1983. Saat ini sebagai Guru Besar di Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung.

Saiful Akbar, memperoleh gelar Sarjana Teknik (S.T.) Jurusan Informatika Institut Teknologi Bandung, lulus tahun 1997. Memperoleh gelar Magister Teknik (M.T.) Program Pasca Sarjana Magister Teknik Informatika Institut Teknologi Bandung, lulus tahun 2002. Memperoleh gelar Doctor of Technology (Dr.techn) Doctoral Program Johannes Kepler Universitat Linz, Austria, lulus tahun 2007. Menyelesaikan Program Paska Doktoral di NTNU, Norwegia, lulus tahun 2010. Saat ini sebagai Dosen Tetap di Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung.

ISSN : 2302-3805